Building an Application Framework for Integrative Genomics

Heta N. Ray, BDS, MS [1, 2], Vamsi K. Mootha, MD [3], Aziz A. Boxwala, MBBS, PhD[1]
[1]Decision Systems Group, Brigham & Women's Hospital, Harvard Medical School, Boston, MA
[2]Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA
[3]Whitehead Institute/MIT Center for Genome Research, Cambridge, MA

The accelerated pace of biological research and the current availability of whole-genome data sets provides significant new sources of functional insight. We designed an architecture and framework for software to query and explore such data in an orderly and iterative fashion. The architecture is intended to provide an extensible platform for developing web based bioinformatics applications and to offer a flexible and end-user-extensible software environment to explore and integrate disparate biological data sources. This will enable the user to explore existing relationships and discover new functional relationships among these data.

INTRODUCTION

Integration of diverse genomic data from cross-platform technologies can spotlight promising candidates with increased confidence, as opposed to relying on a single functional genomics measure, which may suffer from incomplete coverage, imperfect sensitivity, or low specificity [1]. Various specialized algorithms and programs have been developed to identify mapping/associations among various types of concepts (such as DNA, RNA, proteins, and disease). However these applications remain isolated and require manual effort to assemble and integrate information. The manual approach makes searches by users time-consuming and error-prone, and when used repeatedly the searching can be inconsistent.

We designed a framework that makes it possible to integrate these programs into one application. The framework targets mainly two types of users and attempts to provide a unified environment that is useful and reasonably flexible to them: (1) The biologist interested in exploratory data analysis using an interactive user interface, in applying existing data analyses methodologies to new data, in repeating and iterating a set of analysis tasks, and in modifying an analysis methodology to solve a new problem; and (2) The informatician interested in creating an integrated analysis pipeline or a specific application using new and existing analysis programs.

METHODS

The framework allows the user to add interfaces to new data sources (which may be databases, files, or even other programs that dynamically generate associations among concepts) and design new filters in the analysis pipeline. The informatician can extend the application by programming to the well-defined interface. The core framework was implemented using Java classes and Java beans. The web-based user interface is generated using java server pages. An xml file is used to specify the custom application pipeline.

We implemented a demonstration application using data from three sources: (1) Medgene, a web-based program which generates from Medline abstracts, a lists of genes associated with a given disease [2], (2) an RNA microarray dataset from Genomics Institute of the Novartis Research Foundation (San Diego) which profiles the expression of over 10,000 genes, and (3) Genome position annotation tables from the University of California (Santa Cruz) genome browser (http://genome.ucsc.edu/) which is used to select chromosomal regions of interest specified by the user. With these data sources chained in a pipeline, one can find associations between the disease and candidate genes for that disease.

DISCUSSION

We plan to create a user-friendly interface, to assist in the development of custom applications. Although the model seemed adequate for the implemented data sources, we need to validate the generalizability of the model.

Acknowledgments

This research was supported by the National Library of Medicine Training Grant LM07092 and Howard Hughes Medical Institute postdoctoral fellowship.

References

1. Mootha VK, Lepage P, Miller K, Bunkenborg J, Reich M, Hjerrild M, Delmonte T, et al. Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics.. Proc Natl Acad Sci U S A. 2003 Jan 21;100(2):605-10.
2. MedGene, http://hipseq.med.harvard.edu/MEDGENE/about_medgene.html, Accessed 03/06/03